The SyMoGIH project : publishing and sharing historical data on the semantic web

Francesco Beretta, Djamel Ferhod, Séverine Gedzelman, Pierre Vernus

The SyMoGIH project : publishing and sharing historical data on the semantic web

Francesco Beretta francesco.beretta@ish-lyon.cnrs.fr,

with the contribution of

Djamel Ferhod djamel.ferhod@ish-lyon.cnrs.fr,
Séverine Gedzelman severine.gedzelman@ens-lyon.fr,
Pierre Vernus pierre.vernus@ens-lyon.fr

UMR 5190 LARHRA, CNRS / Université de Lyon, France

Abstract

Key-terms : Digital History, Database, Text Encoding Initiative, Linked Data, Ontology for History

1. The SyMoGIH project and linked data

The SyMoGIH project has created an open modular platform for storing geo-historical information. The benefit of the platform is to allow researchers to share their data and texts in a collaborative environment. Up to now, about fifty students and scholars are contributing or contributed to the information collection, and ten research projects are using the platform to store data concerning different domains, such as intellectual, economic, social, institutional or religious history. The richness and heterogeneity of the shared information requires a generic data model which was designed with Merise (ERD) modeling method (cf. *Beretta, Vernus 2012*) and implemented in a relational postgreSQL database, to which users can connect via a user friendly AJAX web application (cf. Beretta, Vernus & Hours 2012). In parallel, we have recently proposed an environment using eXist-db to share XML/TEI encoded texts. The semantic annotation of named entities and knowledge units (i.e. informations) that we find in the texts is achieved by linking the semantic tags to the resources defined in the relational database.

In carrying out the latter, we realized the importance of identifying the database objects using URIs. In addition to the websites we offer for publishing the different datasets related to specific projects (for instance www.patronsdefrance.fr), we also created a generic website to deliver to the public the whole of our authority files and the knowledge units which the participating scholars decided to make public (www.symogih.org). As a first step, this website provides dereferencing of our URIs in form of a web page with the description of the resource (e.g. www.symogih.org/resource/Actr195 for Johannes Kepler) but it raises the issue of providing dereferencement in form of RDF data and, more widely, of connecting our data to those produced elsewhere. For this purpose, we made an ontology that suits our needs.

2. From a project-specific ontology to a generic semantic model for historians

The generic nature of the SyMoGIH data model allows to easily transpose it to an OWL DL ontology. This

project-specific ontology is designed to publish our data and it is based on four main classes : *Object*, *KnowledgeUnit*, *Role* and *Sourcing* (cf. Image 1). An *Object* can be a person, a place, a concept, a bibliographical object, etc. and can be linked to a *KnowledgeUnit* through a *Role*. A superclass *AssociatedObject* gathers Objects and KnowledgeUnits: this allows a KnowledgeUnit to be associated to another one as if it was an object. KnowledgeUnits are divided in two subclasses: an *Information* expresses knowledge as it is constructed by the historian using critical method and extracting it from different sources; a *Content* reproduces knowledge as it was meant by one and only one source, even if we know the source is wrong about an event date or circumstances. In both cases, *Sourcing* provides the origin of each knowledge unit. Specific KnowledgeUnits' and Roles' types appear as instances of the KnowledgeUnitType and RoleType classes: this means that the ontology is versatile, can be gradually expanded and virtually handles any type of knowledge.

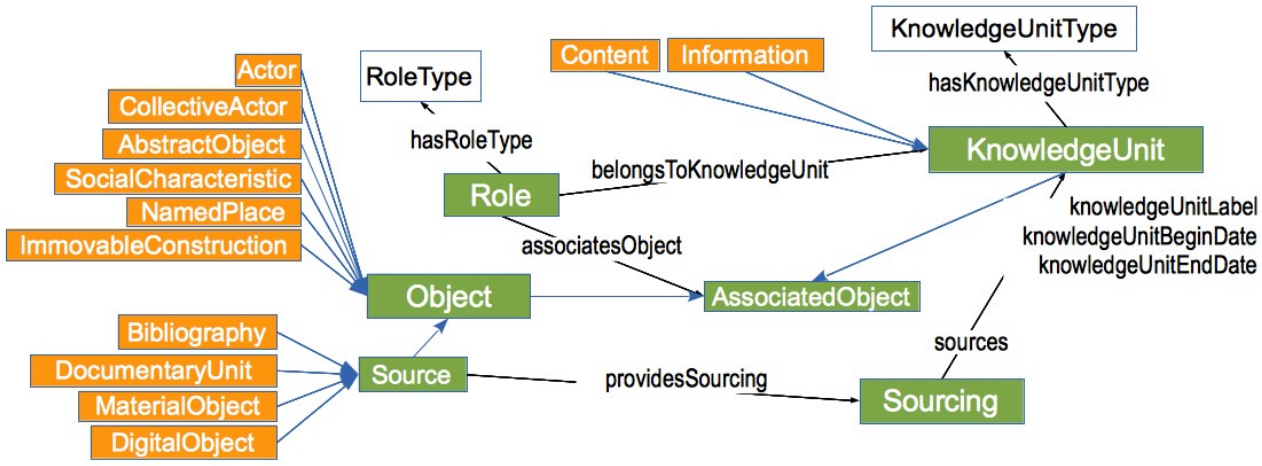

Image 1 : Classes and properties of the Symogih ontology (An instance like “Johannes Kepler (1571-1594)” belongs to the class http://symogih.org/ontology/Actor)

The SyMoGIH ontology bears close resemblance to similar ontologies elaborated by historians and scholars interested in the development of prosopographical databases, particularly the “factoïd” data model developed in King's College Department of Digital Humanities, London (Bradley/Pasin, 2013) and the “Aspekte” data model developed by the Personendaten-Repositorium (http://pdr.bbaw.de/) (Berlin-Brandenburgische Akademie der Wissenschaften). Although independently conceived, the SyMoGIH ontology shows also close similarities to the Simple Event Model ontology (van Hage et al. 2011) and the TemporalEntity in CIDOC-CRM (Doerr, 2003). This convergence of different semantic models treating time-related information has recently risen a discussion between their designers about the definition of a common ontology for publishing and sharing data produced by historical research.

3. Publishing and querying linked data

Pending the achievement of this very important discussion, we used our ontology to create an RDF view on our collaborative database. The implementation uses D2RQ[1] to create a SPARQL endpoint by means of a term map realized according to the RDB2RDF principles[2] that rewrites the database structure according to the SyMoGIH ontology. To increase the SPARQL endpoint performance and allow more sophisticated queries, we periodically dump the available data from D2RQ to an OpenLink Virtuoso triple store[3] and we use OntoWiki[4] for a visual presentation of the dataset. This project is still in experimental phase: at the moment, we are manually interlinking our resources with the one's found in DBPedia, IDRef[5], etc. and we

1 http://d2rq.org/
2 http://www.w3.org/2001/sw/wiki/RDB2RDF
3 http://virtuoso.openlinksw.com/
4 http://aksw.org/Projects/OntoWiki.html
5 http://www.idref.fr/autorites/autorites.html

are testing federated queries to visualize and compare informations coming from different datasets. We are also testing semi-automatic interlinking with other data providers, like DBPedia or GND[6]. This will allow us not only to publish our data on the semantic web but also to compare the quality of available datasets and enrich them, offering new interesting perspectives to researchers in history (cf. *Meroño-Peñuela et al 2013*).

References

Beretta, F., & Vernus, P. (2012). Le projet SyMoGIH et la modélisation de l'information: une opération scientifique au service de l'histoire. Les Carnets du LARHRA, (1), 81–107. (http://halshs.archives-ouvertes.fr/halshs-00677658)

Beretta, F., Vernus, P., & Hours, B. (2012). Le Système modulaire de gestion de l'information historique (SyMoGIH): une plateforme collaborative et cumulative de stockage et d'exploitation de l'information géo-historique. Presented at the Digital Humanities 2012 in Hamburg (http://larhra.ish-lyon.cnrs.fr/iDocuments/SyMoGIH/Poster_SYMOGIH3_jpg.pdf).

Michele Pasin and John Bradley (2013), Factoid-based prosopography and computer ontologies: Towards an integrated approach, Literary and Linguistic Computing Advance Access published June 29, 2013.

Doerr, M. (2003). The CIDOC CRM - An ontological approach to semantic interoperability of metadata. AI Magazine, 24, 2003.

Meroño-Peñuela, A., van Erp, M., Breure, L., Scharnhorst, A., Schlobach, S., & van Harmelen, F. (2013). Semantic Technologies for Historical Research: A Survey. *Semantic Web journal*.

van Hage, W. R., Malaisé, V., Segers, R., Hollink, L., & Schreiber, G. (2011). Design and use of the Simple Event Model (SEM). Web Semantics: Science, Services and Agents on the World Wide Web, 9(2), 128–136. doi:10.1016/j.websem.2011.03.003

Vernus, P. (2009). ANR Système d'information: "Patrons et patronat français - XIXe - XXe siècles" (SIPPAF). Retrieved from http://halshs.archives-ouvertes.fr/halshs-00474109

Links

http://virtuoso.openlinksw.com/
http://www.idref.fr/autorites/autorites.html
http://www.w3.org/2001/sw/rdb2rdf/r2rml/
http://d2rq.org/
http://pdr.bbaw.de/
http://www.patronsdefrance.fr
http://www.loa.istc.cnr.it/DOLCE.html
http://symogih.org/ontology/Actor
http://www.dnb.de/EN/gnd

6 http://www.dnb.de/EN/Standardisierung/GND/gnd_node.html